# TIM245: Data Mining

**Homework #2:** Due Friday, 9 June 2017

**Instructions for Homework # 3:**

* You are allowed to discuss homework problems with other members of course, however, your problem solutions must be **distinctly your own work**, and not a copy of any other student’s work.
* In the assignment you will make use of the following tools: RStudio, and Weka. Before starting the assignment, please download the tools from the links provided below:

1. Jupyter ([http://jupyter.org/)](http://openrefine.org/))
2. R ([https://www.rstudio.com/products/rstudio/)](http://openrefine.org/))

* Please organize your answers to the bolded questions into a well-structured 4-6 page report and submit a hard-copy in class on Monday.

**Problem Statement**

In this homework assignment, you will apply unsupervised learning algorithms to discover interesting patterns in the results of a survey about the opinions of students on a variety of topics ranging from music to spending habits. More specifically you will apply cluster analysis and association analysis to answer the following questions:

* **Cluster Analysis**: what are the general groups of students based on their survey responses?
* **Association Analysis**: are there any relationships (co-occurrence) between the answers to the survey questions?

Before starting the assignment, please download the survey results dataset from the course webpage: <https://tim245-spring17-01.courses.soe.ucsc.edu/>

**Problem 1: Clustering Analysis**

*Note: before starting this assignment it may be useful to review the tutorial* [*https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/*](https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/) *to get a basic understanding of how to create and run notebooks in Jupyter.*

The purpose of this problem is to get experience applying and evaluating the following clustering methods:

1. K-Means
2. DB-Scan
3. Hierarchical Clustering (Agglomerative)
4. Gaussian Mixture Models (Expectation Maximization)

To this end we will find clusters of similar students based on their demographic information and their responses to the survey questions. The survey questions are across the following 7 categories:

1. Music Preferences
2. Movie Preferences
3. Hobbies and Interests
4. Phobias
5. Health Habits
6. Personality Traits, Views on Life, and Opinions
7. Spending Habits

*The survey questions for each category are included in the appendix.*

In order to make the clustering more interpretable, and actionable, we will focus on just one of the seven categories and the provided demographic information.

Please select one category, e.g. music preferences, to focus on for the clustering. Load the survey dataset into the Jupyter Notebook and create a dataset that consists of the selected category and the demographic information:

*import pandas*

*#path is where the file is located , e.g. “/Users/tylermunger/Documents/tim245/hw3/”*

*dataset = pandas.read\_csv('$PATH$/survey\_dataset.csv', sep=',')*

*#remove NAs*

*dataset = dataset.dropna()*

*dataset = dataset.reset\_index(drop=True)*

*#segment based on the survey section*

*music = dataset.iloc[:, 0:19]*

*movies = dataset.iloc[:, 19:31]*

*hobbies = dataset.iloc[:, 31:63]*

*phobias = dataset.iloc[:, 63:73]*

*health = dataset.iloc[:, 73:76]*

*personality = dataset.iloc[:, 76:133]*

*spending = dataset.iloc[:, 133:140]*

*demographics = dataset.iloc[:, 140:150]*

*selected\_category = movies*

*#combine demographics and selected category*

*dataset = pandas.concat([selected\_category, demographics], axis=1)*

*#encode the catorical variables*

*dataset = pandas.get\_dummies(dataset)*

Before clustering the data, it is important to understand of the attributes and instances that are in the dataset. Some basic tools for performing exploratory data analysis in Python are shown below:

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*#print the first 10 rows*

*dataset.head(10)*

*#descriptive statistics*

*dataset.describe()*

*plt.figure(figsize=(20, 10))*

*#subplot syntax is rows, total number of plots, current plot index*

*plt.subplot(3,3,1)*

*plt.scatter(dataset['Age'], dataset['Height'])*

*plt.subplot(3,3,2)*

*plt.hist(dataset['Age'],bins=50)*

*plt.subplot(3,3,3)*

*sns.boxplot(dataset['Age'])*

*plt.show()*

*Note: each notebook cell will only display one output so the descriptive statistics and plots should be run in separate notebook cells. An excellent comprehensive reference for EDA in Python can be found at https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python*

Visually exploring the dataset is an important step of understanding and validating any clustering results. This task becomes especially difficult when working with high-dimensional datasets such as the survey results dataset. One approach for addressing this issue is the T-SNE which maps the higher dimensional space down to two dimensions while trying to preserve the distances between instances.

*from sklearn.manifold import TSNE*

*#create the TSNE representation*

*tsne = TSNE(n\_components=2, verbose=1, perplexity=50, n\_iter=500)*

*dataset\_tsne = tsne.fit\_transform(dataset)*

*#get the values for the two dimensional representation (x,y) for each instance in the dataset*

*tsne\_x = dataset\_tsne [:,0]*

*tsne\_y = dataset\_tsne [:,1]*

*#plot the data*

*scatter = plt.scatter(tsne\_x, tsne\_y, alpha = .75,s = 100)*

*plt.show()*

*Note: an excellent explanation of how to use the T-SNE method can be found at http://distill.pub/2016/misread-tsne/*

Next, we will create and visualize the clustering for each of the four methods:

*from sklearn import cluster*

*from matplotlib import pyplot*

*from sklearn import mixture*

*from sklearn.cluster import AgglomerativeClustering*

*import numpy as np*

*from sklearn import metrics*

*plt.rcParams['image.cmap'] = 'jet'*

*#k is the number of clusters and epsilon is radius in db-scan*

*k = 5*

*epsilon = 5*

*distance\_metric = 'euclidean'*

*#create the clustering models*

*kmeans = cluster.KMeans(n\_clusters=k)*

*dbscan = cluster.DBSCAN(eps= epsilon)*

*hierarchical = cluster.AgglomerativeClustering(n\_clusters=k, affinity=distance\_metric, linkage='ward')*

*gaussian\_mm = mixture.GaussianMixture(n\_components=k)*

*#fit the models to the dataset*

*kmeans.fit(dataset)*

*dbscan.fit(dataset)*

*hierarchical.fit(dataset)*

*gaussian\_mm.fit(dataset)*

*#get the cluster labels for each instance*

*kmeans\_labels = kmeans.labels\_*

*dbscan\_labels = dbscan.labels\_*

*hierarchical\_labels = hierarchical.labels\_*

*gaussian\_mm\_labels = gaussian\_mm.predict(dataset)*

*#compute the silhouette score for each clustering*

*kmeans\_silhouette\_score = metrics.silhouette\_score(dataset, kmeans\_labels, metric=distance\_metric)*

*#must have more than 1 cluster to compute the silhouette\_score for dbscan*

*noise\_index = np.argwhere(dbscan\_labels==-1)*

*if(np.unique(np.delete(dbscan\_labels,noise\_index)).size > 1):*

*dbscan\_silhouette\_score = metrics.silhouette\_score(dataset, dbscan\_labels, metric=distance\_metric)*

*else:*

*dbscan\_silhouette\_score = None*

*hierarchical\_silhouette\_score = metrics.silhouette\_score(dataset, hierarchical\_labels, metric=distance\_metric)*

*gaussian\_mm\_silhouette\_score = metrics.silhouette\_score(dataset, gaussian\_mm\_labels, metric=distance\_metric)*

*#plot the results using the T-SNE representation*

*plt.figure(figsize=(20, 10))*

*plt.subplot(2,2,1)*

*plt.scatter(tsne\_x, tsne\_y, c= kmeans\_labels)*

*plt.title('K Means' + "\n" + 'Silhouette Score: ' + str(kmeans\_silhouette\_score))*

*plt.subplot(2,2,2)*

*plt.scatter(tsne\_x, tsne\_y, c= dbscan\_labels)*

*plt.title('Density Based' + "\n" + 'Silhouette Score: ' + str(dbscan\_silhouette\_score))*

*plt.subplot(2,2,3)*

*plt.scatter(tsne\_x, tsne\_y, c= hierarchical\_labels)*

*plt.title('Hierarchical Clustering' + "\n" + 'Silhouette Score: ' + str(hierarchical\_silhouette\_score))*

*plt.subplot(2,2,4)*

*plt.scatter(tsne\_x, tsne\_y, c= gaussian\_mm\_labels)*

*plt.title('Gaussian Mixture' + "\n" + 'Silhouette Score: ' + str(gaussian\_mm\_silhouette\_score))*

*plt.show()*

*Note: an excellent explanation of clustering methods in sklearn can be found at*

*http://scikit-learn.org/stable/modules/clustering.html*

Compute the clustering results for different values of k and epsilon and plot the results:

*#scores for the parameters*

*kmeans\_silhouette\_scores = []*

*dbscan\_silhouette\_scores = []*

*hierarchical\_silhouette\_scores = []*

*gaussian\_mm\_silhouette\_scores = []*

*distance\_metric = 'euclidean'*

*#loop through different values for k and compute the silhouette score*

*for k in range(2,50,1):*

*kmeans = cluster.KMeans(n\_clusters=k)*

*hierarchical = cluster.AgglomerativeClustering(n\_clusters=k, affinity=distance\_metric, linkage='ward')*

*gaussian\_mm = mixture.GaussianMixture(n\_components=k)*

*kmeans.fit(dataset)*

*hierarchical.fit(dataset)*

*gaussian\_mm.fit(dataset)*

*kmeans\_labels = kmeans.labels\_*

*hierarchical\_labels = hierarchical.labels\_*

*gaussian\_mm\_labels = gaussian\_mm.predict(dataset)*

*kmeans\_silhouette\_score = metrics.silhouette\_score(dataset, kmeans\_labels, metric=distance\_metric)*

*hierarchical\_silhouette\_score = metrics.silhouette\_score(dataset, hierarchical\_labels, metric=distance\_metric)*

*gaussian\_mm\_silhouette\_score = metrics.silhouette\_score(dataset, gaussian\_mm\_labels, metric=distance\_metric)*

*kmeans\_silhouette\_scores.append(kmeans\_silhouette\_score)*

*gaussian\_mm\_silhouette\_scores.append(gaussian\_mm\_silhouette\_score)*

*hierarchical\_silhouette\_scores.append(hierarchical\_silhouette\_score)*

*#loop through different values for epsilon and compute the silhouette score*

*for epsilon in range(2,50,1):*

*dbscan = cluster.DBSCAN(eps=epsilon)*

*dbscan.fit(dataset)*

*dbscan\_labels = dbscan.labels\_*

*#must have more than 1 cluster to compute the silhouette\_score*

*noise\_index = np.argwhere(dbscan\_labels==-1)*

*if(np.unique(np.delete(dbscan\_labels,noise\_index)).size > 1):*

*dbscan\_silhouette\_score = metrics.silhouette\_score(dataset, dbscan\_labels, metric=distance\_metric)*

*else:*

*dbscan\_silhouette\_score = None*

*dbscan\_silhouette\_scores.append(dbscan\_silhouette\_score)*

*#plot silhouette score as function of k and epsilon*

*plt.figure(figsize=(20, 10))*

*plt.subplot(2,4,1)*

*plt.plot(range(2,50,1),kmeans\_silhouette\_scores)*

*plt.title('K-Means')*

*plt.subplot(2,4,2)*

*plt.plot(range(2,50,1),dbscan\_silhouette\_scores)*

*plt.title('DB-Scan Clustering')*

*plt.subplot(2,4,3)*

*plt.plot(range(2,50,1),hierarchical\_silhouette\_scores)*

*plt.title('Hierarchical Clustering')*

*plt.subplot(2,4,4)*

*plt.plot(range(2,50,1),gaussian\_mm\_silhouette\_scores)*

*plt.title('Gaussian Mixture Model')*

*plt.show()*

Lastly, we select one clustering based on the evaluation and save the cluster labels.

*#k is the number of clusters and epsilon is radius in db-scan*

*k = 5*

*epsilon = 5*

*distance\_metric = 'euclidean'*

*#create the clustering models*

*kmeans = cluster.KMeans(n\_clusters=k)*

*dbscan = cluster.DBSCAN(eps= epsilon)*

*hierarchical = cluster.AgglomerativeClustering(n\_clusters=k, affinity=distance\_metric, linkage='ward')*

*gaussian\_mm = mixture.GaussianMixture(n\_components=k)*

*#fit the models to the dataset*

*kmeans.fit(dataset)*

*dbscan.fit(dataset)*

*hierarchical.fit(dataset)*

*gaussian\_mm.fit(dataset)*

*#get the cluster labels for each instance*

*kmeans\_labels = kmeans.labels\_*

*dbscan\_labels = dbscan.labels\_*

*hierarchical\_labels = hierarchical.labels\_*

*gaussian\_mm\_labels = gaussian\_mm.predict(dataset)*

*#get the labels from the selected algorithm, e.g. kmeans*

*selected\_labels = kmean\_labels*

*#add the labels to the dataset*

*labeled\_dataset = dataset*

*labeled\_dataset['cluster\_label'] = pandas.Series(selected\_labels)*

*# path is where you want to save the file, e.g. “/Users/tylermunger/Documents/tim245/hw3/”*

*labeled\_dataset.to\_csv('$PATH$/survey\_dataset\_labeled.csv', sep=',')*

Please answer the following question:

1. **Based on the results, which clustering method do you recommend using for the data-set? Explain why.**
2. **How many clusters did you find in the dataset? How did you select the method parameters, k or epsilon?**
3. **Explore the cluster results using Open-Refine or Excel. Pick 2-3 clusters and try to generalize, i.e. create a persona, for the people (instances) in the cluster.**
4. **Describe how personas from the clustering results could potentially be used?**

**Extra Credit:**

**Experiment with a distance metric other than euclidean. How do the results change?**

**Problem 2: Association Analysis for**

In this problem you will create a set of association rules where and are sets of the answers provided in the survey. Each association rule will therefore describe relationship in people’s opinions and preferences, e.g. people who like rock music and pop music did not like classical music.

We will use the arules package to generate the association rules:

*#install the apriori package*

*install.packages("arules")*

*library(arules)*

*library(arulesViz)*

*#path is where the file is located , e.g. “/Users/tylermunger/Documents/tim245/hw3/”*

*dataset <- read.csv("$PATH#/survey\_dataset.csv ")*

*#convert the dataset to nominal attributes so that it is suitable for rule mining*

*dataset <- data.frame(sapply(dataset, function(x) as.factor(as.character(x))))*

*#n is the number of rules we want to examine*

*n = 10*

*support <- 0.1*

*confidence <- 0.5*

*#generate the rules*

*rules <- apriori(dataset, parameter = list(supp=support, conf=confidence))*

*#a rule is redundant if a more general rules with the same or a higher confidence exists*

*rules.pruned = rules[!is.redundant(rules)]*

*#plot the pruned rules with respect to support, confidence, and lift*

*plot(rules.pruned)*

*#sort by lift and take the top n*

*rules.pruned.sorted <- sort(rules.pruned, by="lift")*

*top\_n\_rules <- head(rules.pruned.sorted, n=n)*

*#print out the top rules*

*inspect(top\_n\_rules)*

*#visualize the top-n rules*

*plot(top\_n\_rules, method="graph", control=list(type="items"))*

Please answer the following questions:

1. **Experiment with different values for support and confidence. How do the discovered patterns change? What threshold do you recommend using for support and confidence?**
2. **Identify 2-5 interesting rules generated using your selected support and confidence threshold. What is the interpretation of the rule? What is underlying rationale or reason for the rule, e.g. the diapers -> beer rule was because young fathers were sent to the store to buy diapers.**
3. **What are some of the potential applications of the generated rules? For example, how could the rules be used for marketing products to students?**

# Appendix: Survey Questions

### **MUSIC PREFERENCES**

1. **I enjoy listening to music.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I prefer.**: Slow paced music 1-2-3-4-5 Fast paced music (integer)
3. **Dance, Disco, Funk**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
4. **Folk music**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
5. **Country**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
6. **Classical**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
7. **Musicals**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
8. **Pop**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
9. **Rock**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
10. **Metal, Hard rock**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
11. **Punk**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
12. **Hip hop, Rap**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
13. **Reggae, Ska**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
14. **Swing, Jazz**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
15. **Rock n Roll**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
16. **Alternative music**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
17. **Latin**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
18. **Techno, Trance**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
19. **Opera**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

### **MOVIE PREFERENCES**

1. **I really enjoy watching movies.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **Horror movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
3. **Thriller movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
4. **Comedies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
5. **Romantic movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
6. **Sci-fi movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
7. **War movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
8. **Tales**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
9. **Cartoons**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
10. **Documentaries**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
11. **Western movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
12. **Action movies**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

### **HOBBIES & INTERESTS**

1. **History**: Not interested 1-2-3-4-5 Very interested (integer)
2. **Psychology**: Not interested 1-2-3-4-5 Very interested (integer)
3. **Politics**: Not interested 1-2-3-4-5 Very interested (integer)
4. **Mathematics**: Not interested 1-2-3-4-5 Very interested (integer)
5. **Physics**: Not interested 1-2-3-4-5 Very interested (integer)
6. **Internet**: Not interested 1-2-3-4-5 Very interested (integer)
7. **PC Software, Hardware**: Not interested 1-2-3-4-5 Very interested (integer)
8. **Economy, Management**: Not interested 1-2-3-4-5 Very interested (integer)
9. **Biology**: Not interested 1-2-3-4-5 Very interested (integer)
10. **Chemistry**: Not interested 1-2-3-4-5 Very interested (integer)
11. **Poetry reading**: Not interested 1-2-3-4-5 Very interested (integer)
12. **Geography**: Not interested 1-2-3-4-5 Very interested (integer)
13. **Foreign languages**: Not interested 1-2-3-4-5 Very interested (integer)
14. **Medicine**: Not interested 1-2-3-4-5 Very interested (integer)
15. **Law**: Not interested 1-2-3-4-5 Very interested (integer)
16. **Cars**: Not interested 1-2-3-4-5 Very interested (integer)
17. **Art**: Not interested 1-2-3-4-5 Very interested (integer)
18. **Religion**: Not interested 1-2-3-4-5 Very interested (integer)
19. **Outdoor activities**: Not interested 1-2-3-4-5 Very interested (integer)
20. **Dancing**: Not interested 1-2-3-4-5 Very interested (integer)
21. **Playing musical instruments**: Not interested 1-2-3-4-5 Very interested (integer)
22. **Poetry writing**: Not interested 1-2-3-4-5 Very interested (integer)
23. **Sport and leisure activities**: Not interested 1-2-3-4-5 Very interested (integer)
24. **Sport at competitive level**: Not interested 1-2-3-4-5 Very interested (integer)
25. **Gardening**: Not interested 1-2-3-4-5 Very interested (integer)
26. **Celebrity lifestyle**: Not interested 1-2-3-4-5 Very interested (integer)
27. **Shopping**: Not interested 1-2-3-4-5 Very interested (integer)
28. **Science and technology**: Not interested 1-2-3-4-5 Very interested (integer)
29. **Theatre**: Not interested 1-2-3-4-5 Very interested (integer)
30. **Socializing**: Not interested 1-2-3-4-5 Very interested (integer)
31. **Adrenaline sports**: Not interested 1-2-3-4-5 Very interested (integer)
32. **Pets**: Not interested 1-2-3-4-5 Very interested (integer)

### **PHOBIAS**

1. **Flying**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
2. **Thunder, lightning**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
3. **Darkness**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
4. **Heights**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
5. **Spiders**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
6. **Snakes**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
7. **Rats, mice**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
8. **Ageing**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
9. **Dangerous dogs**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
10. **Public speaking**: Not afraid at all 1-2-3-4-5 Very afraid of (integer)

### **HEALTH HABITS**

1. **Smoking habits**: Never smoked - Tried smoking - Former smoker - Current smoker (categorical)
2. **Drinking**: Never - Social drinker - Drink a lot (categorical)
3. **I live a very healthy lifestyle.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

### **PERSONALITY TRAITS, VIEWS ON LIFE & OPINIONS**

1. **I take notice of what goes on around me.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I try to do tasks as soon as possible and not leave them until last minute.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. **I always make a list so I don't forget anything.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. **I often study or work even in my spare time.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. **I look at things from all different angles before I go ahead.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. **I believe that bad people will suffer one day and good people will be rewarded.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. **I am reliable at work and always complete all tasks given to me.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
8. **I always keep my promises.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
9. **I can fall for someone very quickly and then completely lose interest.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
10. **I would rather have lots of friends than lots of money.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
11. **I always try to be the funniest one.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
12. **I can be two faced sometimes.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
13. **I damaged things in the past when angry.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
14. **I take my time to make decisions.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
15. **I always try to vote in elections.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
16. **I often think about and regret the decisions I make.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
17. **I can tell if people listen to me or not when I talk to them.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
18. **I am a hypochondriac.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
19. **I am emphatetic person.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
20. **I eat because I have to. I don't enjoy food and eat as fast as I can.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
21. **I try to give as much as I can to other people at Christmas.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
22. **I don't like seeing animals suffering.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
23. **I look after things I have borrowed from others.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
24. **I feel lonely in life.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
25. **I used to cheat at school.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
26. **I worry about my health.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
27. **I wish I could change the past because of the things I have done.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
28. **I believe in God.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
29. **I always have good dreams.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
30. **I always give to charity.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
31. **I have lots of friends.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
32. **Timekeeping.**: I am often early. - I am always on time. - I am often running late. (categorical)
33. **Do you lie to others?**: Never. - Only to avoid hurting someone. - Sometimes. - Everytime it suits me. (categorical)
34. **I am very patient.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
35. **I can quickly adapt to a new environment.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
36. **My moods change quickly.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
37. **I am well mannered and I look after my appearance.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
38. **I enjoy meeting new people.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
39. **I always let other people know about my achievements.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
40. **I think carefully before answering any important letters.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
41. **I enjoy childrens' company.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
42. **I am not afraid to give my opinion if I feel strongly about something.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
43. **I can get angry very easily.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
44. **I always make sure I connect with the right people.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
45. **I have to be well prepared before public speaking.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
46. **I will find a fault in myself if people don't like me.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
47. **I cry when I feel down or things don't go the right way.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
48. **I am 100% happy with my life.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
49. **I am always full of life and energy.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
50. **I prefer big dangerous dogs to smaller, calmer dogs.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
51. **I believe all my personality traits are positive.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
52. **If I find something the doesn't belong to me I will hand it in.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
53. **I find it very difficult to get up in the morning.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
54. **I have many different hobbies and interests.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
55. **I always listen to my parents' advice.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
56. **I enjoy taking part in surveys.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
57. **How much time do you spend online?**: No time at all - Less than an hour a day - Few hours a day - Most of the day (categorical)

### **SPENDING HABITS**

1. **I save all the money I can.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I enjoy going to large shopping centres.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. **I prefer branded clothing to non branded.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. **I spend a lot of money on partying and socializing.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. **I spend a lot of money on my appearance.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. **I spend a lot of money on gadgets.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. **I will hapilly pay more money for good, quality or healthy food.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

### **DEMOGRAPHICS**

1. **Age**: (integer)
2. **Height**: (integer)
3. **Weight**: (integer)
4. **How many siblings do you have?**: (integer)
5. **Gender**: Female - Male (categorical)
6. **I am**: Left handed - Right handed (categorical)
7. **Highest education achieved**: Currently a Primary school pupil - Primary school - Secondary school - College/Bachelor degree (categorical)
8. **I am the only child**: No - Yes (categorical)
9. **I spent most of my childhood in a**: City - village (categorical)
10. **I lived most of my childhood in a**: house/bungalow - block of flats (categorical)